SPECTRAL-ENVELOPE-BASED LEAST SIGNIFICANT BIT MANAGEMENT FOR LOW-DELAY BIT-ERROR-ROBUST SPEECH CODING

Ryosuke Sugiura, Yutaka Kamamoto, Takehiro Moriya

NTT Communication Science Labs., Nippon Telegraph and Telephone Corp.,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

We have devised a method for bit assignment of quantized frequency spectra aiming at its use in low-delay bit-error-robust speech compression. The proposed method, least significant bit management (LSBM), controls the least significant bits of the spectra based on their envelope to make them represented by fixed bit rates, which guarantees the range of the damage caused by the bit error which makes the mismatch of the spectral envelopes between the encoder and the decoder. In addition, we relate the method to the linear predictive coding scheme and show its performance and robustness in a speech codec by objective and subjective evaluations. The codec based on this method, having bit-error robustness with only 1.5-ms algorithmic delay, can be useful at such situations as real-time speech communication with non-IP protocols.

***Index Terms*—** Speech coding, bit assignment, spectral envelope, low delay, bit error

1. INTRODUCTION

The development and spread of Internet protocol (IP) have made a large influence on the trend of speech compression. Especially, the concept of packet freed us from the risks of bit errors, allowing us to use powerful variable-length coding tools such as arithmetic coding even in real-time speech communication. These tools have enhanced the performance of many codecs and contributed to the state-of-the-art speech codec standards like 3GPP Enhanced voice services (EVS) codec [1].

However, there still remains a demand on speech communication without IP in conditions where bit errors will occur. Among various Internet of things (IoT) systems, although they are basically expected to depend on IP, some use cases may need local communication or observation, which require total optimization with more simple protocol having smaller overhead at the expense of bit-error correction capacity. As a well-known conventional example, wireless microphones need low-delay codecs dealing with bit errors. Using variable-length coding tools in these cases, bit errors will make the decoder mistake the sample length to decode, with its influence lasting for a long period and collapsing the whole communication at the worst case. This possibility of the mistakes will not be zero even if the code is protected by some error-correction codes. For this reason, Adaptive differential pulse code modulation (ADPCM) [2] and its derivatives like Enhanced APT-X [3], codecs used for compressing high-quality speech in the above conditions, have fixed bit assignments and tend to use high bit rate, which limits the number of devices working simultaneously at limited communication bandwidth. In addition, the high bit rate makes the bit-error problem harder, even for low bit error rates (BERs), compared with the case of low-bit-rate radio communication.

Therefore, this work sets the goal at designing an efficient low-delay and bit-error-robust codecs with lower bit rates than the conventional ones to transmit more sounds simultaneously with high sound quality. To deal not only with speech but also music items, we focus here on a modified-discrete-cosine-transform (MDCT)-based codec, which can easily control the quantization noise, instead of codecs based on strict speech models, for example, ITU-T Rec. G.728, G.729 and Pitch synchronous innovation code excited linear prediction (PSI-CELP) [4–6] although they are robust against bit errors and some of them have contributed to non-IP mobile communication.

Guaranteeing the decoded frame-wise sample length under bit-error conditions requires strictly fixing the total bits used for each frame. At higher bit rates, vector quantization, one of the well-used fixed-length coding techniques, have difficulties in avoiding its exponentially increasing computation costs although there are many extensions to reduce them [7–9]. Therefore, codecs working at the bit rates focused here are more preferable to use sample-wise bit assignment with scalar quantization. In this case, the compression performance of the codec greatly depends on how to assign the bits to each frequency spectra and how to represent the bit assignment itself.

In this paper, we thus introduce an error-robust realization of bit assignment, Least significant bit management (LSBM), from the perspective of spectral envelopes. This method is designed to reduce the effects of the bit errors occurred in the bit assignment on the degradation of the reconstructed frequency spectra. The paper first quickly reviews the conventional idea of bit assignment for frequency spectra, providing the relationship with the spectral envelopes, and describes the

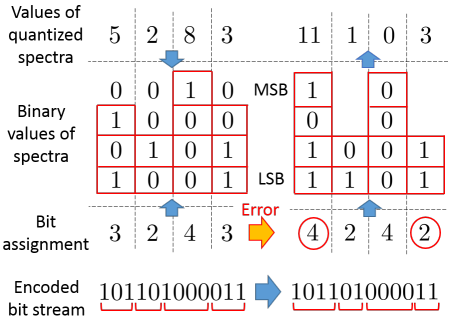

Fig. 1. Simple coding of quantized spectra with bits assigned to each frequency. Bit assignment indicates the number of the red blocks used for packing each spectrum into the bit stream. If the bit error, in the yellow arrow, changes the information of the bit assignment as in the red circles, the bit steam read by the decoder will be shifted, resulting in large errors even in the frequency where the bit assignment is correct.

method of LSBM. Then, the integration of LSBM into the codec is shown, followed by some objective and subjective evaluations.

2. ENVELOPE-BASED BIT ASSIGNMENT

Many methods of bit assignment for frequency spectra are based on the idea of adaptive transform coding in [10], which insists on the optimality of bit assignment in the sense of rate-distortion tradeoffs. The following discussions focuses on assigning bits to positive values by omitting polarity of the spectra, which can be coded independently or combined with the positive values by folding and interleaving as in [11]. Assuming real-valued MDCT spectra $\{X_n \in [0, \infty)\}_{n=0}^{N-1}$, each distributed according to independent probability density functions (pdfs) which have a same shape but different variances $\{\sigma_n^2\}_{n=0}^{N-1}$, the probability distribution and the optimal expected code length $\{b_n\}_{n=0}^{N-1}$ of each spectra scalar-quantized by quantization steps $\{d_n\}_{n=0}^{N-1}$, can be approximately written with a pdf $p(X)$ of variance 1 as $\left\{\frac{d_n}{\sigma_n} p\left(\frac{d_n X}{\sigma_n}\right)\right\}_{n=0}^{N-1}$ and

$$\begin{aligned} b_n &\simeq \int_0^\infty -\frac{d_n}{\sigma_n} p\left(\frac{d_n X}{\sigma_n}\right) \log_2 \left(\frac{d_n}{\sigma_n} p\left(\frac{d_n X}{\sigma_n}\right)\right) \mathrm{d}X \\ &= \log_2 \frac{\sigma_n}{d_n} + C(p(X)), \end{aligned} \tag{1}$$

respectively, where $C(p(X))$ is a constant depending only on the shape $p(X)$ of the distributions. If the quantization steps are small enough, the expected energy of the quantization noise will proportionate d_n^2, and thus the rate-distortion optimizing problem of a given bit rate B becomes

$$\min_{\{d_n\}} \sum_{n=0}^{N-1} d_n^2 \ \text{s.t.}\ B = \sum_{n=0}^{N-1} b_n. \tag{2}$$

This leads to the optimal quantization steps

$$d_n = 2^{-\frac{B}{N}} \left(\prod_{j=0}^{N-1} \sigma_j\right)^{1/N}, \tag{3}$$

uniform among the frequencies n, and the bit assignment

$$b_n = \log_2 H_n + C(p(X)), \ \ H_n \equiv \frac{\sigma_n}{\left(\prod_{j=0}^{N-1} \sigma_j\right)^{1/N}}, \tag{4}$$

of which differences among the frequencies n depend only on the normalized standard deviation $\{H_n\}_{n=0}^{N-1}$.

This normalized standard deviation is often approximated by an all-pole model parameterized as, for example, linear prediction (LP) coefficients or line spectrum pairs (LSP) [12–16] since the spectral envelope derived from the model has a property that its geometric average among the frequencies becomes 1, just as $\{H_n\}_{n=0}^{N-1}$ has. Therefore, the encoder can optimally code the spectra by the combination of scalar quantization and bit assignment based on the parameterized spectral envelope. The decoder reconstructs the spectra by first decoding the envelope and then reading the bit stream according to the bit assignment the envelope suggests. More concretely speaking, parameterizing the bits to be assigned are equivalent to parameterizing the logarithmic spectral envelope.

As mentioned in the introduction, we have to use strictly fixed bit assignment, namely, the bit assignment $\{b_n\}_{n=0}^{N-1}$ should be integer as is the case of the codec in [10]. However, in cases where the encoder simply packs the code of spectra sequentially into the bit stream as in Fig. 1, even a little errors in the bit assignment disturb the whole order of the bits read by the decoder, making unexpected decoding errors. Indeed, it may be effective to modify the scanning order of the bit planes like in [17], but it still has a risk of unexpected collapse.

In the next section, we show that the encoding and decoding processes can be more robust against the errors in the bit assignment, or in the spectral envelope, by using the fact that it can be divided into the constant part independent on frequencies and the frequency-wise differences which sum up to zero and can be approximated by the spectral envelopes.

3. LEAST SIGNIFICANT BIT MANAGEMENT BASED ON SPECTRAL ENVELOPE

3.1. Least significant bit management

As mentioned above, assuming infinite-support independent identical-shape distributions, the optimal bit assignment can be divided into the constant term and the envelope-dependent terms, the terms related to $\{H_n\}_{n=0}^{N-1}$, regardless of the distributions we use. In fact, this assumption practically holds at finite-support signals. That is to say, if the spectral envelopes are precise enough, each spectrum divided by its corresponding envelope are expected to fit within a constant bit length. However, quantizing after dividing the spectra, the energy of

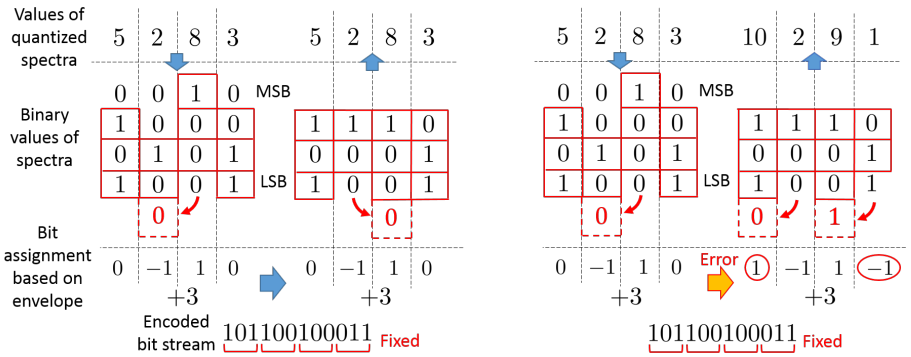

Fig. 2. Coding of quantized spectra with LSBM. If the bit assignment is sent correctly, the decoder can reconstruct the quantized spectra correctly (left half). Even if errors change the bit assignment as in the red circles, the order of the bit stream read by the decoder will not change (right half).

the quantization noise becomes non-uniform among the frequencies and fails to minimize eq. (2).

Here, we can use the property of the logarithmic values of the spectral envelopes summing up to zero, which enables us to realize invertible 'division' in the integer region by managing the LSBs of the quantized values of the spectra. The proposed method, LSBM, performs the bit assignment as in Fig. 2, to spectra $\{5, 2, 8, 3\}$ for example. The bit assignment is first divided into the constant term, say 3 bits, and the envelope-dependent terms aggregating zero, say $\{0, -1, 1, 0\}$ bits. The encoder then takes off the LSBs from the spectra according to their corresponding envelope-dependent terms if they are positive: 1 bit taken off from 8. If they are negative, the encoder sets the LSBs taken off to the LSBs of the spectra according to them: 1 bit set to 2. Since the total of the envelope-dependent terms are zero, the processed spectra can be represented by fixed-length codes according to the constant term: $\{5, 4, 4, 3\}$ can be written within 3 bits. Predetermining the rules of the order for taking and setting the LSBs, the decoder can reconstruct the correct quantized spectra by reversing the processes of the encoder.

In the LSBM scheme, the decoder reads out the bit stream by fixed bit-length intervals so that even if the bit assignment changed by errors, the influence of the errors do not spread to the whole order of the bits, which guarantees that the most significant bits (MSBs) of the spectra will be preserved where their corresponding envelope-dependent terms are correct. With respect to the example in Fig. 2, 3 bits of MSBs of the spectra are guaranteed in this sense. Therefore, there are less chances where the errors crucially distort the outline of the spectra as in Fig. 1, and the decoder can avoid large perceptual degradation.

3.2. Spectral-envelope coding at low-delay conditions

Of course, the spectral envelop $\{H_n\}_{n=0}^{N-1}$ can be coded using all-pole models represented by LP coefficients or LSPs. However, since the envelope-dependent terms of the bit assignment need only the integer precision for the LSBM scheme, in low-delay conditions where the frame length and the envelope length are short, it may be efficient to code the envelope-dependent terms $\{\log_2 H_n\}_{n=0}^{N-1}$ directly by vector quantization (VQ), for instance. In these cases, the distortion can be minimized by preparing the patterns of envelope-dependent terms $\{\log_2 H_n^{(m)}\}_{n=0}^{N-1}$ for $j = 0, \cdots, M-1$, which are organized to sum up to zero in each pattern, and searching the pattern minimizing the spectral energy $\sum_{n=0}^{N-1} \left| X_n / H_n^{(m)} \right|^2$, with bit-shift operations, for example.

In fact, taking into account that $\sum_n \log_2 H_n = 0$ for every patterns, this minimization problem is rewritten as

$$\min_m \sum_{n=0}^{N-1} \left(\left| X_n / H_n^{(m)} \right|^2 - \ln \left| X_n / H_n^{(m)} \right|^2 - 1 \right), \qquad (5)$$

in other words, equivalent to the Itakura-Saito (IS) divergence minimization, which LP analysis also deals [18]. Therefore, the vector quantization mentioned above have the same characteristics in the sense of envelope fitting in LP.

4. APPLICATION TO SPEECH CODEC

We designed a simple low-delay super-wide-band speech codec using the proposed scheme. The codec compresses 32-kHz sampling rate 16-bit depth monaural signals into around 96 kbps, working for 32- or 64-sample frames with quarter overlaps, which makes 1.5- or 3-ms algorithmic delay. It first transforms the input signals into MDCT spectra and performs perceptual weighting by self-determined block companding (SDBC) [19]. Then, the envelope-dependent terms of the bit assignment are determined by vector quantization mentioned

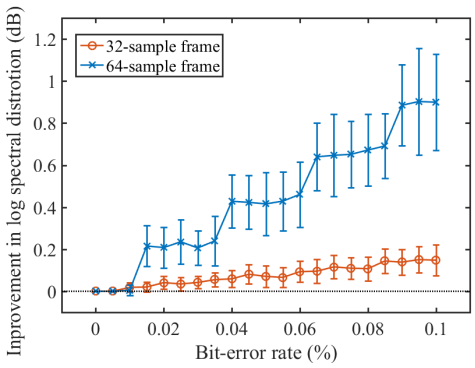

Fig. 3. Item-wise improvement in log spectral distortion for each BER. Item-wise average and standard deviation. Circles and crosses indicate 32- and 64-sample-frame conditions, respectively.

above based on the weighted spectra. The weighted spectra are scalar quantized by a single quantization step, followed by LSBM. The envelope-dependent terms used for LSBM are represented in 4 bits at 32-sample-frame mode and 12 bits at 64-sample frame mode by the vector quantization. Aiming at reconstructing speech signals, the decoder performs pitch enhancement [20] by a post-filter using only the decoded signals.

5. EVALUATION OF ESTIMATION METHOD

At first, to check the robustness of the proposed method, we compared the log spectral distortion [21] of the reconstructed MDCT spectra by artificially making bit errors on the encoded bit streams. Two types of codecs were prepared: one used proposed LSBM for bit assignment and the other used conventional simple sequential bit assignment. We encoded and decoded by these codecs 84 items, about 8 seconds each of speech and audio with 32-kHz sampling rate and 16-bit depth. Among the encoded bit streams, the bits representing the spectral envelopes were randomly altered according to each BERs. Note that the commercial use in this condition usually requires under around 0.01% BER for safety.

Fig. 3 shows the improvement, the difference of the proposed method from the conventional one, in log spectral distortion at each frame-length conditions. The difference between the proposed and conventional bit assignment became clearer as the BER raised. However, the preliminary experiment on the reproduction signal-to-noise ratio (SNR) showed little difference between the conditions. This means the small values such as spectra in higher frequencies were damaged by bit errors more at the conventional bit assignment, and the errors should be more annoying for listeners compared to them in the case of using LSBM.

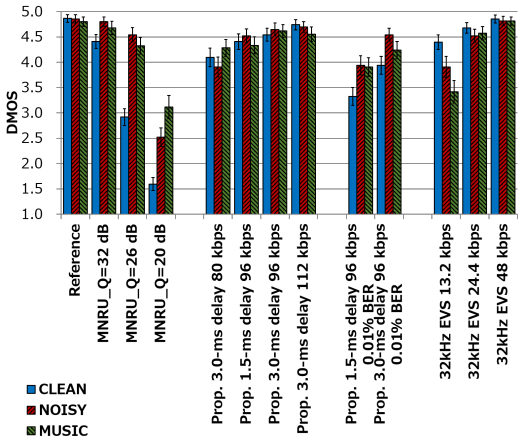

Fig. 4. Condition-wise DMOS of ITU-T Rec. P.800. Error bars indicate 95 % confidence intervals.

6. EVALUATION OF SPEECH CODEC

To evaluate the quality of the proposed codec, we held a subjective experiment for degradation category rating (DCR) based on ITU-T Rec. P.800 [22]. 24 participants rated, according to five-point degradation, four items each of clean speech, noisy speech and music for the respective conditions. For a reference, we prepared EVS codec, known to perform very high quality at mobile communication using more than ten times longer algorithmic delay (32 ms) compared to the proposed coder. Some conditions had 0.01% BER with errors randomly generated based on uniform probability.

Fig. 4 describes the item-wise degradation mean opinion score (DMOS) based on the average over the 96 votes for each condition. Due to the limitations of space, we omitted from the graph the results of Modulated noise reference units (MNRUs) at 14- and 8-dB signal-to-modulated-noise ratio. It can be seen that the proposed codec at 96 kbps displayed high DMOS around 4.5 in average. In addition, the DMOS of the proposed codec with 0.01% bit error was around 4, showing little degradation especially at noisy conditions.

7. CONCLUSION

We presented a bit assignment method for frequency spectra based on addition and subtraction of LSBs according to the spectral envelopes. The proposed method, LSBM, enables to guarantee the range of the damage by the bit errors occurring in the codes for spectral envelopes, which may reduce the annoying noise given by the errors.

Comparison with other bit assignment methods remains for future works. Additionally, the rules may be modified in LSBM for the order of adding and subtracting LSBs to further strengthen its robustness.

8. REFERENCES

[1] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, Lei Miao, Zhe Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, "Overview of the EVS codec architecture," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5698–5702.

[2] K. Sayood, *Introduction to Data Compression*, chapter 11, Elsevier, 2006.

[3] K. Campbell, "Deploying large scale audio over ip networks," in *AES 126th Convention, #7652*, 2009.

[4] ITU-T Rec. G. 728, "Coding of speech at 16 kbit/s using low-delay code excited linear prediction," 2013.

[5] ITU-T Rec. G. 729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," 1996.

[6] S. Miki, K. Mano, T. Moriya, K. Oguchi, and H. Ohmuro, "A pitch synchronous innovation celp (psi-celp) coder for 2-4kbit/s," *Proc. ICASSP 1994*, 4 1994.

[7] Biing-Hwang Juang and A. Gray, "Multiple stage vector quantization for speech coding," in *ICASSP 1982*, May 1982, vol. 7, pp. 597–600.

[8] T. Moriya, "Two-channel conjugate vector quantizer for noisy channel speech coding," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 866–874, Jun 1992.

[9] N. Iwakami, T. Moriya, and S. Miki, "High-quality audio-coding at less than 64 kbit/s by using transform-domain weighted interleave vector quantization (TwinVQ)," in *ICASSP 1995*, May 1995, vol. 5, pp. 3095–3098.

[10] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 299–309, Aug 1977.

[11] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Trans. on Information Theory*, vol. 46, no. 1, pp. 121–135, Jan. 2000.

[12] G. Fuchs, C.R. Helmrich, G. Markovic, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5723–5727.

[13] T. Backstrom and C.R. Helmrich, "Arithmetic coding of speech and audio spectra using tcx based on linear predictive spectral envelopes," in *Proc. ICASSP 2015*, Apr. 2015, pp. 5127–5131.

[14] R. Sugiura, Y. Kamamoto, N. Harada, H. Kameoka, and T. Moriya, "Resolution warped spectral representation for low-delay and low-bit-rate audio coder," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 2, pp. 288–299, 2015.

[15] R. Sugiura, Y. Kamamoto, N. Harada, H. Kameoka, and T. Moriya, "Optimal coding of generalized-gaussian-distributed frequency spectra for low-delay audio coder with powered all-pole spectrum estimation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 8, pp. 1309–1321, Aug. 2015.

[16] T. Backstrom, *Speech coding*, Springer, 2017.

[17] T. Li, S. Rahardja, and S. N. Koh, "Perceptually prioritized bit-plane coding for high-definition advanced audio coding," in *Eighth IEEE International Symposium on Multimedia*, Dec 2006, pp. 245–252.

[18] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun.*, vol. 53-A, no. 1, pp. 36–43, 1970.

[19] R. Sugiura, Y. Kamamoto, N. Harada, T. Kawanishi, and T. Moriya, "CLEAR: Conditionally Lossless Encoding under Allowed Rates for Low-Delay Sound Data Transmission," in *AES 143th Convention, #9899*, Oct 2017.

[20] D. Malah and R. Cox, "A generalized comb filtering technique for speech enhancement," in *ICASSP 1982*, May 1982, vol. 7, pp. 160–163.

[21] W. Kleijn and K. Paliwal, *Speech coding and synthesis*, pp. 433–466, Elsevier, 1995.

[22] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," ITU-T, 1992.